ERIC

John Greene, Ph.D., SRA International
(ERIC Program and Systems Update)

&

Guy Plunkett, Ph.D., U. Wisconsin-Madison
(Annotation/Curation Progress)

BRC - 3
February 6, 2006

Enteropathogen Resource Integration Center

- ERIC focuses on integration of data from 5 enteropathogens, via a pathogen-centric Web portal:
  - Diarrheagenic *E. coli*
  - *Shigella* spp.
  - *Salmonella* spp.
  - *Yersinia enterocolitica*
  - *Yersinia pestis*
- Partnership between personnel at the Genome Center of Wisconsin and SRA International, Rockville MD
- Goals adjusted – not just tools and DBs, but limited analysis…

http://www.ericbrc.org

- ERIC is in continuous and accelerating development
- Pathogen-centric, not technology-centric vision of system
- Community annotation via ASAP was functional on Day 1, and continues to be functional. Portal up March 5, 2005 - no unplanned downtime since then.
- Suggestions, comments, criticism are welcome by e-mail at info@ericbrc.org

Enteropathogen Resource Integration Center

Bioinformatics Resource Center

Welcome

Welcome to ERIC! - the **E**nteropathogen **R**esource **I**ntegration **C**enter.

ERIC is one of eight Bioinformatics Resource Centers (BRC) for Biodefense and Emerging/Re-Emerging Infectious Diseases. Funded by the National Institute of Allergy and Infectious Diseases (NIAID), ERIC serves as an information resource for five members of the bacterial family Enterobacteriaceae.

Developed by SRA International, Inc. and the Genome Center of Wisconsin, ERIC is a Web portal intended to integrate many types of biological information for these five organisms and to facilitate research into therapeutics, diagnostics, and vaccines. ERIC provides data and tools to support investigator-driven data analysis. Current and evolving tools* include:

- Community Genome Annotation (ASAP)
  - Find information about the pathogen genomes and sequences
  - Contribute your annotations to a curated database
- Comparative Genomics*
- Microarray Analysis*
- Antigen Prediction*
- Genome Polymorphisms*
- Pathways/Subsystems*
- Text Mining of Abstracts and Biomedical Literature*
- Proteomics*
- Phenotypic Data (including pathogenicity and virulence)*

- ***Diarrheagenic E. coli***
- ***Shigella***
- ***Salmonella***
- ***Yersinia enterocolitica***
- ***Yersinia pestis***

Further Information

To learn more about ERIC and NIAID's Bioinformatics Resource Centers please visit the following links:

- BRC-Central.org
- The SRA/GCW Team and Our Approach
- NIAID Biodefense Web site
- NIAID BRC Program Web site

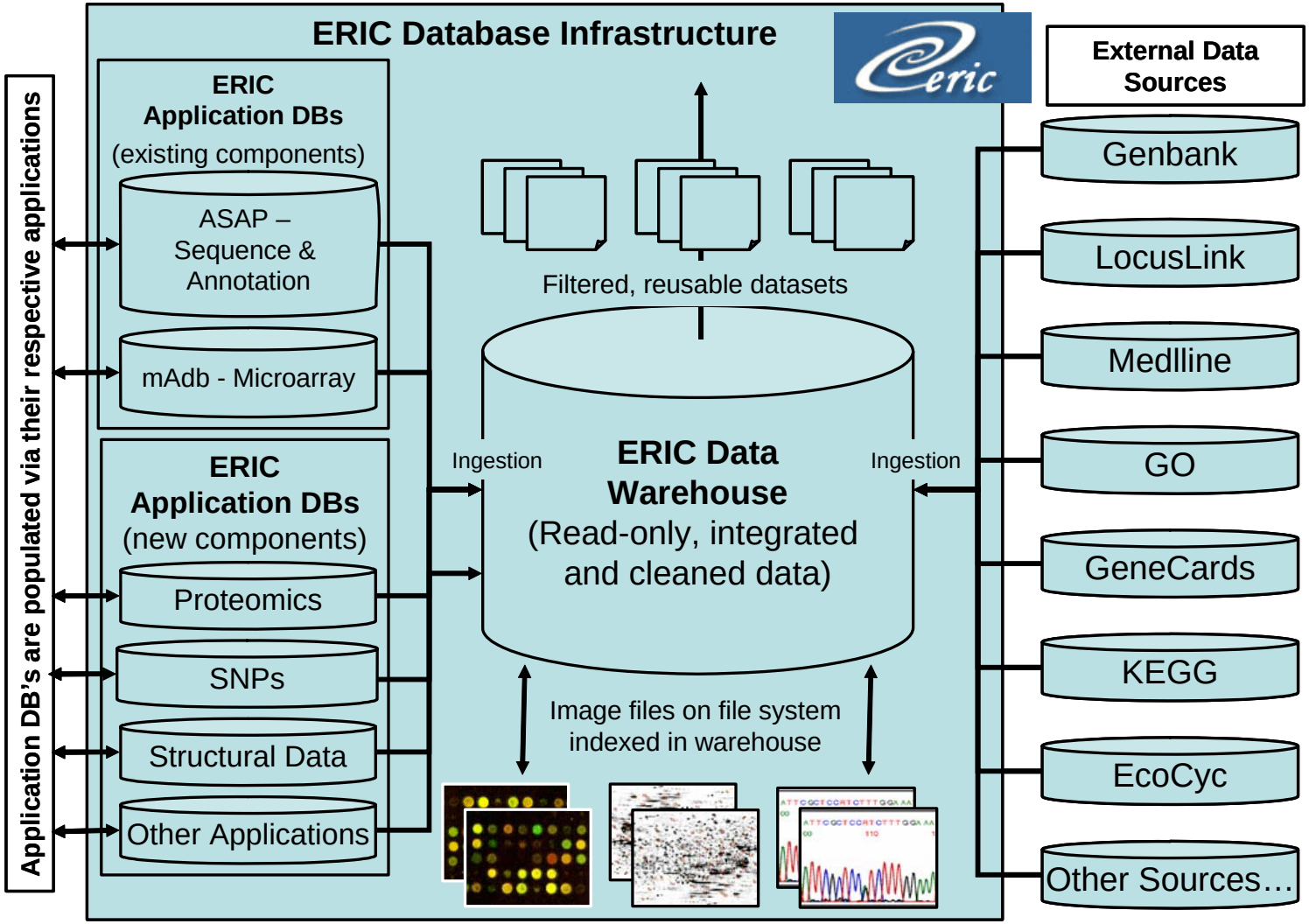
ERIC Announcements

ERIC BRC Has A New Look:
In January 2006 the ERIC BRC portal released a new structure to allow for simpler, more pathogen-centric navigation. Let us know what you think of our new look by e-mailing us at info@ericbrc.org.

Eric Help

For more information on ERIC, please e-mail info@ericbrc.org.

Send questions regarding technical support to administrator@ericbrc.org.



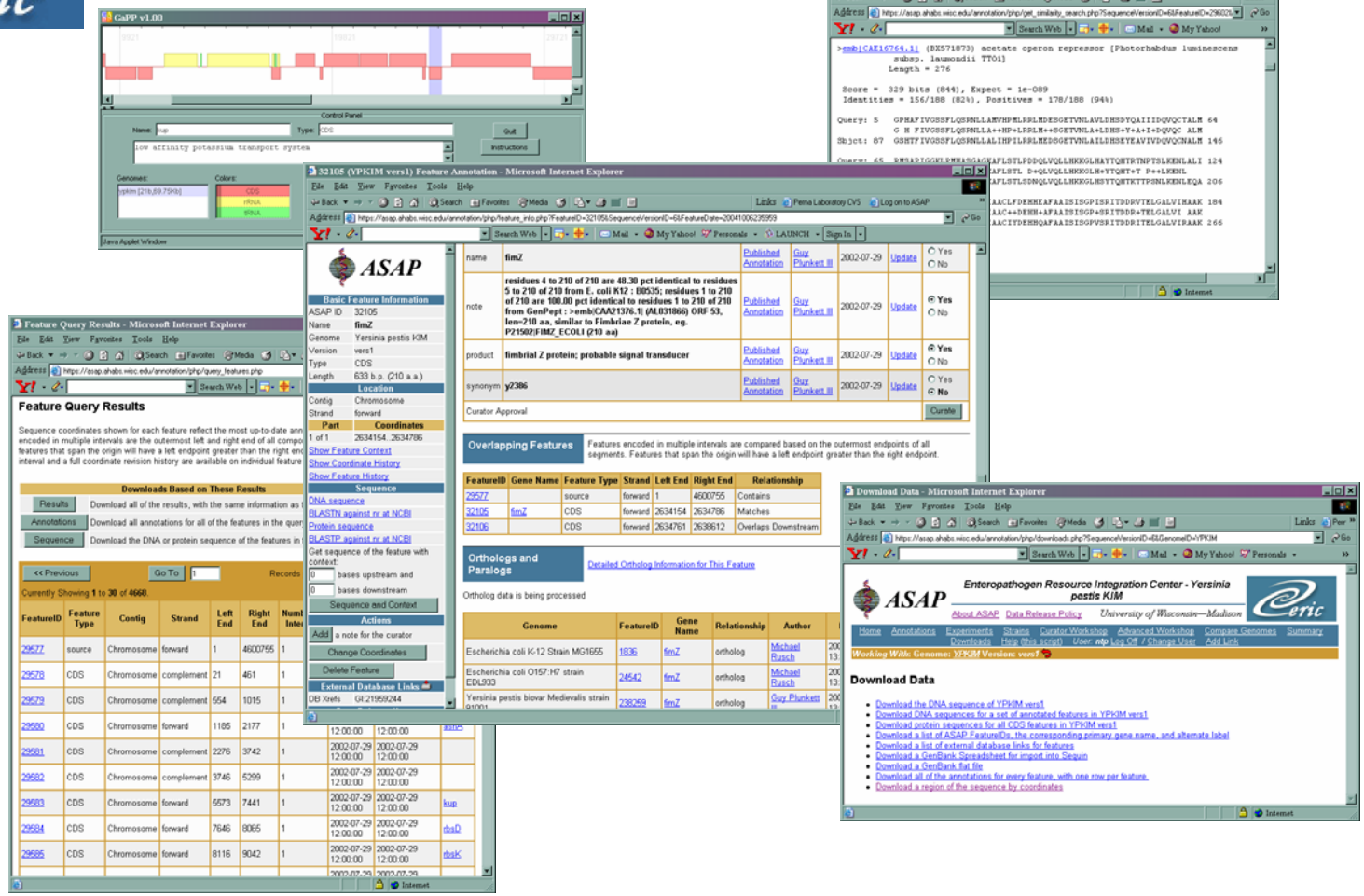
ERIC is beginning to make use of a **data warehouse** approach to store both contributed pathogen data (annotations, sequence, microarray, proteomics, etc.) and data from external sources. We are beginning with integration of microarray data from our mAdb component with annotations from ASAP.

Development Methodologies

- Agile development
  - “Short-cycle Development Iterations
  - Subversion for Configuration Management
  - Bugzilla for bug and change tracking
  - Peer review of software; formal QA and testing
  - Weekly Configuration Control Board (CCB) meetings to review and approve software change requests (SCRs), with participation from our UW teammates.
  - Trying for close interactions with the enterobacterial research community, soliciting feedback on system development and features…will discuss this tomorrow AM
- Focus on delivering scientist-friendly user interfaces that work together

System Components –"big rocks first"

- **ASAP**
  - MAUVE integrated into ASAP at UW
  - Contains voluntary annotations for reference genomes (E.coli K12, uropathogenic E. coli, Yersinia pseudotuberculosis)
  - Synchronization 3x/day with ERIC-ASAP
  - In beta testing round 3
- **mAdb** – for microarray data
  - Scalable microarray system with advanced analysis tools
  - Porting from Sybase to Oracle complete
  - MIAME compliant in version 1.1
  - Testing beginning…
  - Question – what microarray scanners other than Affy, GenePix, QuantArray?
- **GBrowse**
  - Excellent Web integration and very customizable user interface.
  - Integrating with ASAP
  - GBrowse has been adapted by a number of the BRCs and therefore will promote interoperability.
- **JBoss Portal**
  - Just completed switch from Apache Jetspeed to JBoss Portal
  - Allows JSR-168 Portlet compatibility
  - The Alfresco open source content management system – coming very soon
  - Portal design advantage is that users will be able to **customize** it for their particular needs – and change it as their needs change



ASAP meets needs for direct, community-wide input (with authorship and annotation history tracking), multiple annotations of features, evidence codes, using controlled vocabularies, curatorial review, support of cross-genome comparisons, and web-based updating and access. In addition, ASAP links annotations to features, and not specifically to genome assemblies.

2 ASAP Instances

- ERIC contains up-to-date community annotations of ERIC bacterial genomes as well as related genomes useful for comparative genomics.
- 28 ERIC genomes hosted on ERIC-ASAP; 6 reference genomes related to ERIC hosted on UW-ASAP when in production.
- Automated, bi-directional synchronization between UW ASAP and ERIC ASAP instances
- Pre-built MAUVE alignments are already publicly available for a) four *E. coli*, b) two *Shigella*, c) four *Yersinia*, and d) five *Salmonella* genomes.

Increased Computational Needs

- Our most serious current challenge is a computational bottleneck. The number of genomes in ERIC has risen much faster than expected.

- Although the concept of annotating orthologous genes (and gene families via our new EnteroFams), may allow us to keep pace with the genome numbers, we are having a simpler problem merely running computationally intensive bioinformatics analysis tools (e.g. BLAST, InterProScan, HMMer) against the increasing number of these genomes.

- We are examining several methods to increase computational capacity:
  - increasing the size of our computational cluster;
  - making use of BLAST results from BRCs with larger computational resources;
  - obtaining time on other clusters such as the NSF Teragrid and the NIH Biowulf;
  - modifying our analysis SOP and/or algorithm parameters .

Coming….

- Comparative Genomics
- Antigen Prediction
- Genome Polymorphisms
- Pathways/Subsystems
- Proteomics
- Phenotypic Data (including pathogenicity and virulence)

SRA Text Mining Technology

- SRA is an industry leader in natural language processing (NLP)-based text mining
  - Dedicated group of linguists and software engineers
  - Have participated in numerous Government text mining evaluations (e.g. Message Understanding Conference (MUC); Automated Content Extractions (ACE), run by NIST, where NetOwl has been the top scorer in entity and relationship extraction for the past three years.
- Extensive experience in multilingual information extraction, text clustering, and text summarization – this is not keyword searching
  - Numerous commercial and Government clients/applications
  - Financial services and intelligence applications
- Discovering heretofore unrecognized relationships in the literature may be key to designing experiments which will lead to the identification of targets for vaccines, therapeutics, and diagnostics

SRA Text Mining – NetOwl Extractor

- **NetOwl™ Text Mining** - 100 randomly selected abstracts from the ERIC ASAP subsystem were manually annotated with SGML to identify common terms. Overall, this resulted in an average of 45 annotations for each abstract, and a total of 48 different annotated concepts. From this list of 48 concepts, the top ten were prioritized for the creation of Natural Language Processing (NLP) patterns for their extraction from unstructured text. The current prioritized concepts are:

  Gene; gene product; organism and strain; operon; pathogenesis; literature reference; mutant; protein (divided into enzyme and non-enzyme products); encoding (transcription and translation); and regulation.

- NLP patterns were developed by a senior computational linguist working with one of ERIC's bioinformatics scientists, using the original 100 marked up abstracts as a training set. These descriptors recognize the concepts listed above in unstructured text. Known terms for each concept continue to be captured from public sources and used to enhance the effectiveness of recognition. NetOwl Extractor was run on the training set and scores indicating the relevant accuracy and the false positive rate were determined. Existing taxonomies were obtained and reviewed to be used to augment patterns.

- The final step will be validating these patterns on a larger blind set of annotated abstracts to be used to test, configure, and optimize the NetOwl Extractor engine. Creation and markup of this blind test set is beginning.

Enteropathogen Resource Integration Center (ERIC) Scientific Working Group, 2005		
Name	**Affiliation**	**Research Interests**
Emilio Garcia, Ph.D.	Lawrence Livermore National Lab	Leader of *Y. pseudotuberculosis* genome project; *Yersinia* comparative genomics.
Fiona Brinkman, Ph.D.	Simon Fraser U., Canada	Research Director, Bioinformatics, for the Genome Canada Pathogenomics Project; Coordinator of the *Pseudomonas aeruginosa* Community Annotation Project and *Pseudomonas* Genome Database. Expertise in genome analysis, bioinformatics, and systems biology.
James Kaper, Ph.D.	U. Maryland, Baltimore	Leader in the fields of *E. coli* (EPEC, EHEC) pathogenicity, genomics; epidemiology, and vaccine development.
Julian Davies, Ph.D.	U. British Columbia, Canada	World recognized expertise in microbial resistance, microbial ecology, and bacterial diversity.
Robert Perry, Ph.D.	U. Kentucky, Lexington	Dr. Perry is a leading expert on the *Y. pestis* genome, its pathogenicity, and genetics.
Shelley Payne, Ph.D.	U. Texas, Austin	*Shigella* expertise; Expert on iron-regulated pathogenicity in *Shigella* and *Vibrio.*
Stanley Maloy, Ph.D.	San Diego State University	Recognized expert in *Salmonella* genomics, host specificity, comparative genomics of *Salmonella* , expression regulation, and pathogenic mechanisms.
Stephen Calderwood, M.D.	Chief, Infectious Disease Unit, Mass. General Hospital and Professor of Medicine, Harvard Medical School.	Genome variability, evolution; pathogenicity (*E. coli* , *Vibrio*); vaccine development.
Thomas Cebula, Ph.D.	FDA; Director, Office of Applied Research and Safety Assessment	Genomic variability and mutations in *E. coli* and *Salmonella* ; safety and compliance; antibiotic resistance; epidemiology.
Will Gilbert, Ph.D.	U. New Hampshire	Dr. Gilbert was formerly Director of Computing at the Whitehead Institute, and has consulted for or worked at AstraZeneca, Millennium, Genetics Institute, Genome Therapeutics, and Human Genome Sciences. He is the Hamel Professor of Innovation and Technology at UNH, and directs the Bioinformatics Group in the UNH Genome Center. He has been instrumental in developing numerous large-scale genomic databases and analysis tools.

- ERIC has held 3 SWG meetings (2 in-person; one telecon)
- Intend to add a member from industry

Desiderata…

- Will discuss ERIC's community outreach tomorrow…
- Collaborating with NMPDR – ERIC will adopt subsystems concept; NMPDR has expressed interest in mAdb for microarray data.
- IPP distributed as template to other BRCs

ERIC Security and ATO

- ERIC is **fully** compliant with Article H.6 of the contract.
- Our Information Technology Systems Security Plan (ITSSP) and Certification and Assurance (C&A) Package was submitted to and reviewed by NIAID,
- Complies with the requirements of the SOW; the Federal Information Security Management Act of 2002 (FISMA); the Computer Security Act of 1987; Office of Management and Budget (OMB) Circular A-130, Appendix III; and the DHHS Information Security Program Handbook.
- The required official Authority to Operate (ATO) was issued by NIAID's CIO in November 2005, with **r**e-inspection required every 3 years.

ERIC Team:

Scientific Co-Directors:

Eric “Half A Bee” Greene, Ph.D. - Principal Investigator
Erica Perna, Ph.D. - Scientific Co-Director
Eric Blattner, Ph.D. - Scientific Co-Director

ERIC Curators:

Eric Plunkett III, Ph.D. - Senior Curator; Erica Burland, Ph.D.; Eric “the real Eric” Cabot, Ph.D.; Eric Anderson, Ph.D., Eric “the other real Eric” Neeno-Eckwall, Ph.D., Eric Glasner, Ph.D.; Eric Yu Qiu; Eric Mau, Ph.D.

ERIC Technical Team:

- Eric Shaker, Project Manager; Eric Hampton, Tech Lead and Architect; Eric Martell; Erica Shaull; Erica Shetty; Erica Wong; Eric Pot, Ph.D.; Eric Sfeir; Eric Backus
- Eric Liss; Eric Rusch

ERIC Team:

Scientific Co-Directors:

John Greene, Ph.D. - Principal Investigator
Nicole Perna, Ph.D. - Scientific Co-Director
Fred Blattner, Ph.D. - Scientific Co-Director

ERIC Curators:

Guy Plunkett III, Ph.D. - Senior Curator; Val Burland, Ph.D.; Eric Cabot, Ph.D.; Brad Anderson, Ph.D., Eric Neeno-Eckwall, Ph.D., Jeremy Glasner, Ph.D.; Yu Qiu; Bob Mau, Ph.D.

ERIC Technical Team:

- Matt Shaker, Project Manager; Tom Hampton, Tech Lead and Architect; Robin Martell; Lorie Shaull; Panna Shetty; Mary Wong; David Pot, Ph.D.; Rob Sfeir; Mark Backus
- Paul Liss; Michael Rusch

Ladies and Gentlemen…

Guy Plunkett!

Enteropathogen Resource Integration Center

New ERIC and Reference Genomes added since May 2005

3 completed *Shigella* genomes [Chinese Ministry of Science and Technology]

- *S. sonnei* str. 046 (chromosome + 1 plasmid)
- *S. dysenteriae* serotype 1 str. 197 (chromosome + 1 plasmid)
- *S. boydii* serotype 4 str. 227 (chromosome + 1 plasmid)

8 draft genomes (WGS assemblies) [NIAID Microbial Sequencing Center, TIGR]

- *Shigella boydii* serotype 18 str. BS512: 79 contigs
- † *Escherichia coli* str. B7A (ETEC, serotype O148:H28) --198 contigs
- † *Escherichia coli* str. E22 (EPEC, serotype O103:H2) -- 109 contigs
- † *Escherichia coli* str. E110019 (EPEC, serotype O111:H9) -- 115 contigs
- † *Escherichia coli* str. B171 (EPEC, serotype O111:NM) -- 159 contigs
- *Escherichia coli* str. E24377A (ETEC, serotype O139:H28) -- 1 contig (+ plasmids)
- *Escherichia coli* str. 53638 (EIEC, serotype O144) -- 119 contigs
- *Yersinia pestis* str. Angola (biovar Antiqua) -- 121 contigs

3 draft reference genomes (WGS assemblies) [NIAID Microbial Sequencing Center, TIGR]

- † *Escherichia coli* str. F11 (ExPEC) -- 88 contigs
- *Escherichia coli* str. HS (human commensal) -- 1 contig
- *Yersinia pseudotuberculosis* str. IP 31758 -- 1 contig

† these genomes will be remain at draft stage; the others will be taken to closure

Enteropathogen Resource Integration Center

Molecular applications for identifying microbial pathogens in the post-9/11 era

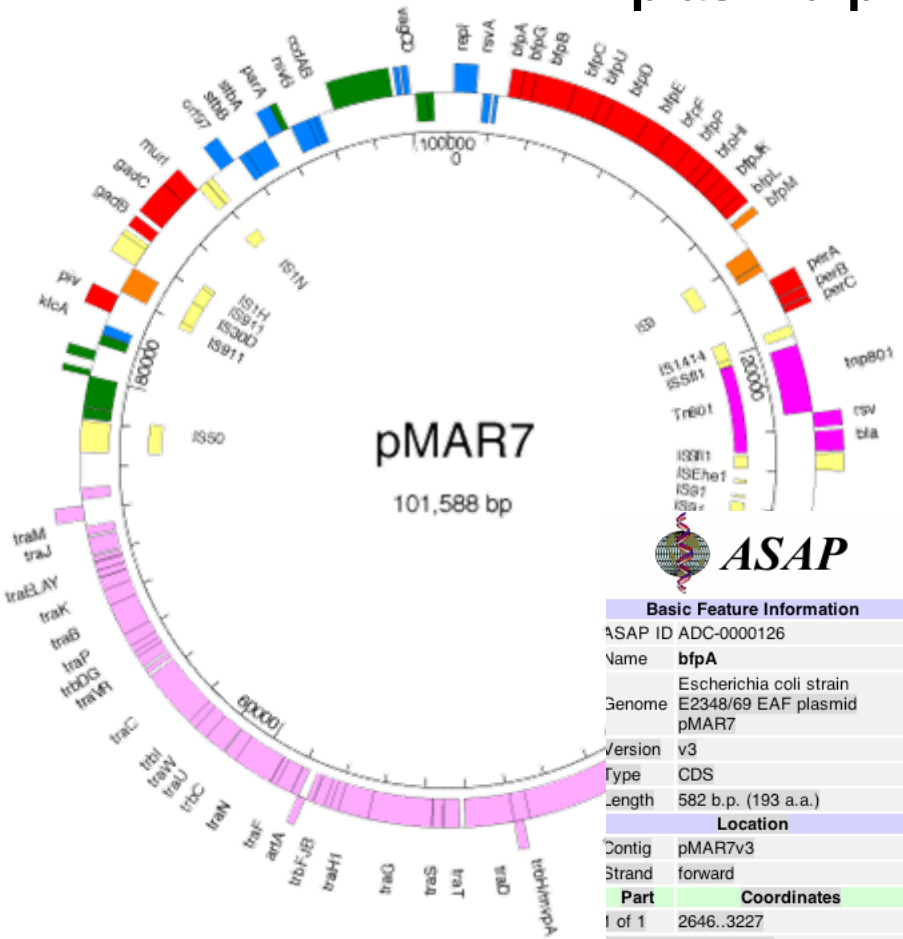
Thomas A Cebula†, Eric W Brown, Scott A Jackson, Mark K Mammel, Amit Mukherjee and J Eugene LeClerc

Genome Sequence Correction: *E. coli* O157:H7 strain EDL933

Four sequence corrections (3 single-base changes and a 12 bp indel) reported by Dr. Tom Cebula's group have been examined and accepted. In addition, ambiguities in the consensus sequence are being examined and it seems that many can be resolved with the existing sequence data. 20% of the genome has been addressed, and 15 single-base indels and about 250 ambiguities have been resolved. Similar data is available for the O157:H7 Sakai strain. These ambiguities are false polymorphisms, obstructing the identification of real polymorphisms (i.e. SNPs) among different EHEC strains.

Enteropathogen Resource Integration Center

The Concept of a Private Genome: Annotation of plasmid pMAR7 from EPEC strain E2348/69



In collaboration with Dr. Jim Kaper, ERIC curator Dr. Val Burland has annotated this plasmid in ASAP as a private genome. Upon publication, it will immediately be made publicly available via ERIC with full annotations and experimental evidence.

A Systematic Annotation Package for Community Analysis of Genomes

About ASAP | Data Release Policy | *University of Wisconsin—Madison*

Home | Annotations | Experiments | Strains | Curator Workshop | Advanced Workshop | Compare Genomes
Uploads | Summary | Downloads | Help (this script) | *User:* ***gp3*** Log Off / Change User | Add Link

Working With: Genome: ***pMAR7*** Version: ***v3***

Skip down to Annotations, Overlapping Features

Basic Feature Information	
ASAP ID	ADC-0000126
Name	**bfpA**
Genome	Escherichia coli strain E2348/69 EAF plasmid pMAR7
Version	v3
Type	CDS
Length	582 b.p. (193 a.a.)
Location	
Contig	pMAR7v3
Strand	forward
Part	**Coordinates**
1 of 1	2646..3227

Show Feature Context
Show Coordinate History
Show Feature History

Sequence

DNA sequence
BLASTN against nr at NCBI
Protein sequence
BLASTP against nr at NCBI
Get sequence of the feature with

Annotations

(view NCBI definitions and examples), (view information on MultiFun)

Type	Data	Evidence	Annotator	Date Annotated	Curator Approval
name	**bfpA**	Experimental - this species or strain	Val Burland	2005-08-26	Uncurated

Overlapping Features

Features encoded in multiple intervals are compared based on the outermost endpoints of all segments. Features that span the origin will have a left endpoint greater than the right endpoint.

FeatureID	Gene Name	Feature Type	Strand	Left End	Right End	Relationship
ADC-0000126	bfpA	CDS	forward	2646	3227	Matches

Enteropathogen Resource Integration Center

Genomes currently being annotated and curated in ERIC ASAP

Diarrheagenic *Escherichia coli*	2	Enterohemorrhagic *E. coli* (EHEC) EDL933 and Sakai
	3	Enteropathogenic *E. coli* (EPEC) E22, B171, and E110019
	2	Enterotoxigenic *E. coli* (ETEC) B7A and E24377A
	1	Enteroinvasive *E. coli* (EIEC) 53638
Shigella spp.	2	*Shigella flexneri* 2457T and 301
	1	*Shigella sonnei* 046
	1	*Shigella dysenteriae* 197
	2	*Shigella boydii* BS512 and 227
Salmonella spp.	2	*Salmonella* Typhi Ty2 and CT18
	1	*Salmonella* Typhimurium LT2
	1	*Salmonella* Paratyphi A 9150
	1	*Salmonella* Choleraesuis B67
Yersinia pestis	2	*Yersinia pestis* biovar Mediaevalis KIM and 91001
	1	*Yersinia pestis* biovar Orientalis CO92
	1	*Yersinia* pestis biovar Antiqua Angola
“orphan” plasmids	2	*Shigella flexneri* invasion plasmids pWR100 and pWR501
	1	*Salmonella* Typhimurium drug resistance plasmid R27
	2	Enteropathogenic *E. coli* plasmids pB171 and pMAR7
reference genomes	6	4 non-diarrheagenic *E. coli* and 2 *Y. pseudotuberculosis*

Enteropathogen Resource Integration Center

Annotation & Curation: by the Numbers

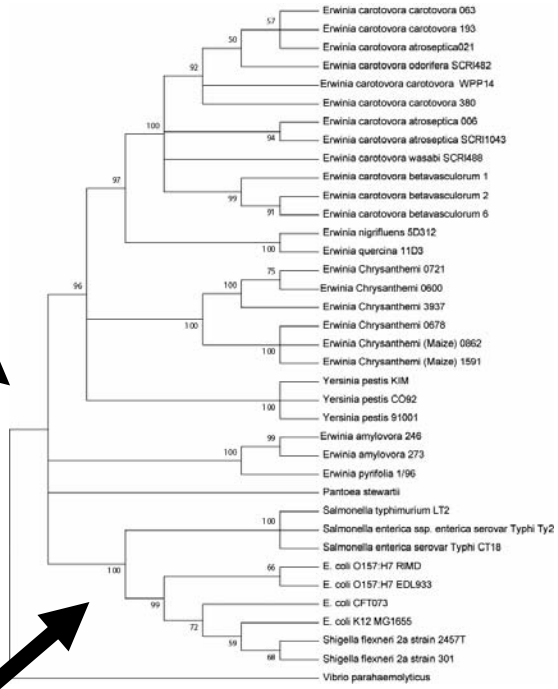
Date	Features Added	Features Annotated	Annotations Added	Annotations Curated	Orthologs Added	Orthologs Curated
(previously)	91171	154241	558417	558417	268505	9072
Sept 2005	5307	5670	41389	26739	531	4446
Oct 2005	124	419	1413	15590	41266	1315
Nov 2005	37732	39468	295784	290156	770	7328
Dec 2005	10161	10299	76717	79717	125	758
Subtotal	53324	55856	415303	412202	42692	13847
Total	144495	210097	973720	970619	311197	22919

Enteropathogen Resource Integration Center

EnteroFams

HMMs representing conserved proteins in Enterobacteriaceae

Roughly 2100 proteins are likely ancestral to all Enterobacteriaceae



Additional proteins are found in taxonomic subsets

Objective: **Create a set of reference protein models to reliably detect members of these conserved families in newly sequenced genomes.**

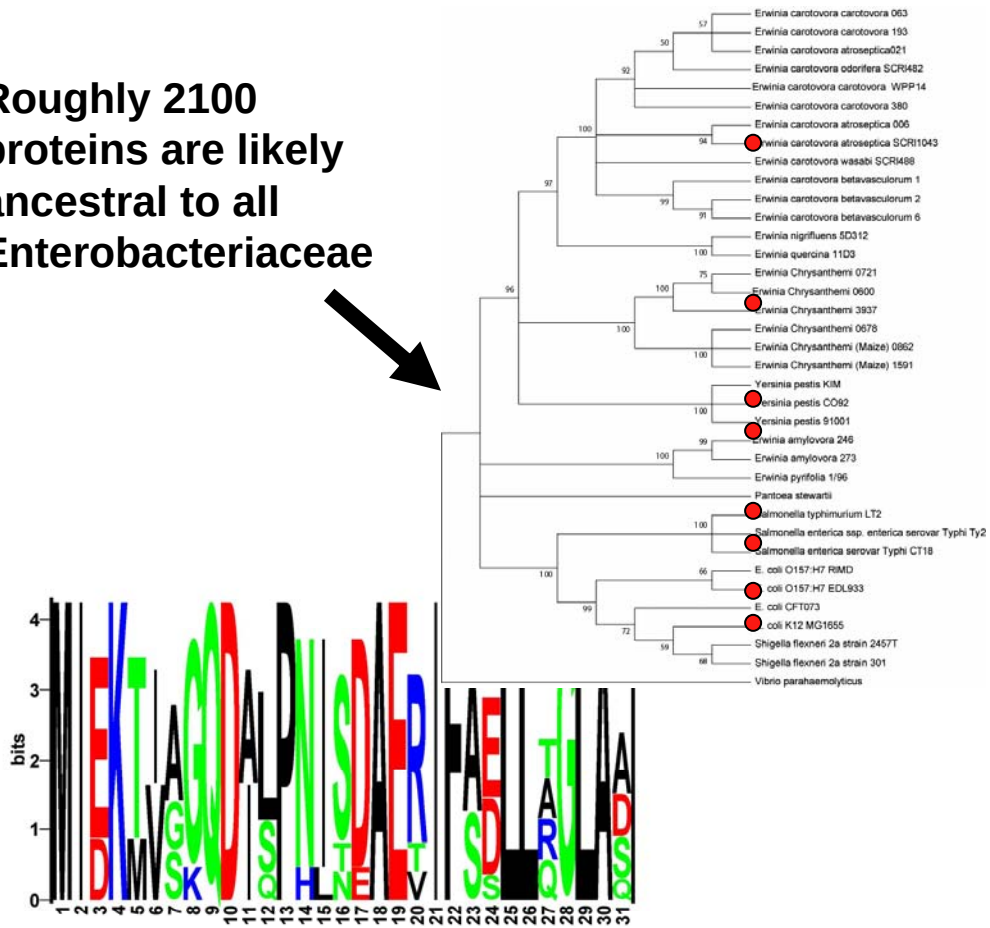
Approach: **Build multiple alignments and HMM profiles based on representative taxa. Validate and establish trusted score thresholds by searching test genomes. Apply trusted thresholds to new genomes.**

Enteropathogen Resource Integration Center

EnteroFams

Round 1: Proteins expected to occur in all Enterobacteriaceae

Roughly 2100 proteins are likely ancestral to all Enterobacteriaceae



1678 sets of putative orthologs used to build HMMs.

1629 families validated.

A trusted score threshold can be established which returns only one match, the putative ortholog, in each test genome

23 genomes searched to identify additional family members

EnteroFam0000008: Trusted threshold = 2308.9

Enteropathogen Resource Integration Center

EnteroFams

Use to improve annotation efficiency and standardization

Genome	Annotation Type	Annotation Value	Evidence Code	Evidence Type	Evidence Value
MG1655	EC number	1.1.1.3	PA	DBN	GenProtEC
MG1655	EC number	2.7.2.4	PA	GBAN	U00096
MG1655	function	enzyme; Amino acid biosynthesis: Threonine	PA	GBAN	U00096
MG1655	GO biological process	GO:0009090	OM	SDO	MultiFun to GO
MG1655	GO biological process	GO:0009088	OM	SDO	MultiFun to GO
MG1655	GO biological process	GO:0009086	OM	SDO	MultiFun to GO
MG1655	GO cellular component	GO:0005737	OM	SDO	MultiFun to GO
MG1655	MultiFun	1.5.1.8	PA	MFDBN	MultiFun
MG1655	MultiFun	7.1	PA	MFDBN	MultiFun
MG1655	MultiFun	1.5.1.21	PA	MFDBN	MultiFun
MG1655	MultiFun	1.5.1.9	PA	MFDBN	MultiFun
MG1655	name	thrA	PA	GBAN	U00096
MG1655	product	bifunctional: aspartokinase I (N-terminal); homoserine dehydrogenase I (C-terminal)	PA	DBN	GenProtEC

EnteroFams added to UW-ASAP as a “Virtual Genome”.

Create a nonredundant set by combining and collapsing annotations from all members.

Curators can augment and edit annotations appropriate to the entire family by editing the EnteroFam annotations.

Propagate EnteroFam annotations to all members.

Update EnteroFam.

Comparative Genomics: Genomes, Genes and Pathogenesis

Gene **acquisition**, **loss**, and **inactivation** are the main mechanisms that contribute to the evolution of a pathogen's fitness.

- Genes can be **acquired** via horizontal transfer -- mediated by plasmids, bacteriophages, and other mobile genetic elements -- giving rise to so-called “pathogenicity islands” and other strain-specific loci.
- Genes can be **lost** by excision of, and exchange among, these same mobile elements.
- Genes can be **inactivated** by the standard gamut of mutations, giving rise to what have come to be called **pseudogenes**.

Enteropathogen Resource Integration Center

Microbial Pseudogenes

• **Mutant** genes in a given genome, defined by comparison to a related genome where the wild-type or "**ancestral**" state is seen.
• **Distinguished** from **missense** mutations, where the gene is still intact but may have altered functionality.
• **Inactivation** by in-frame stop codons, frameshifts, insertion of IS elements, prophages, or islands; also gene remnants, the aftermath of deletions, rearrangements, etc.

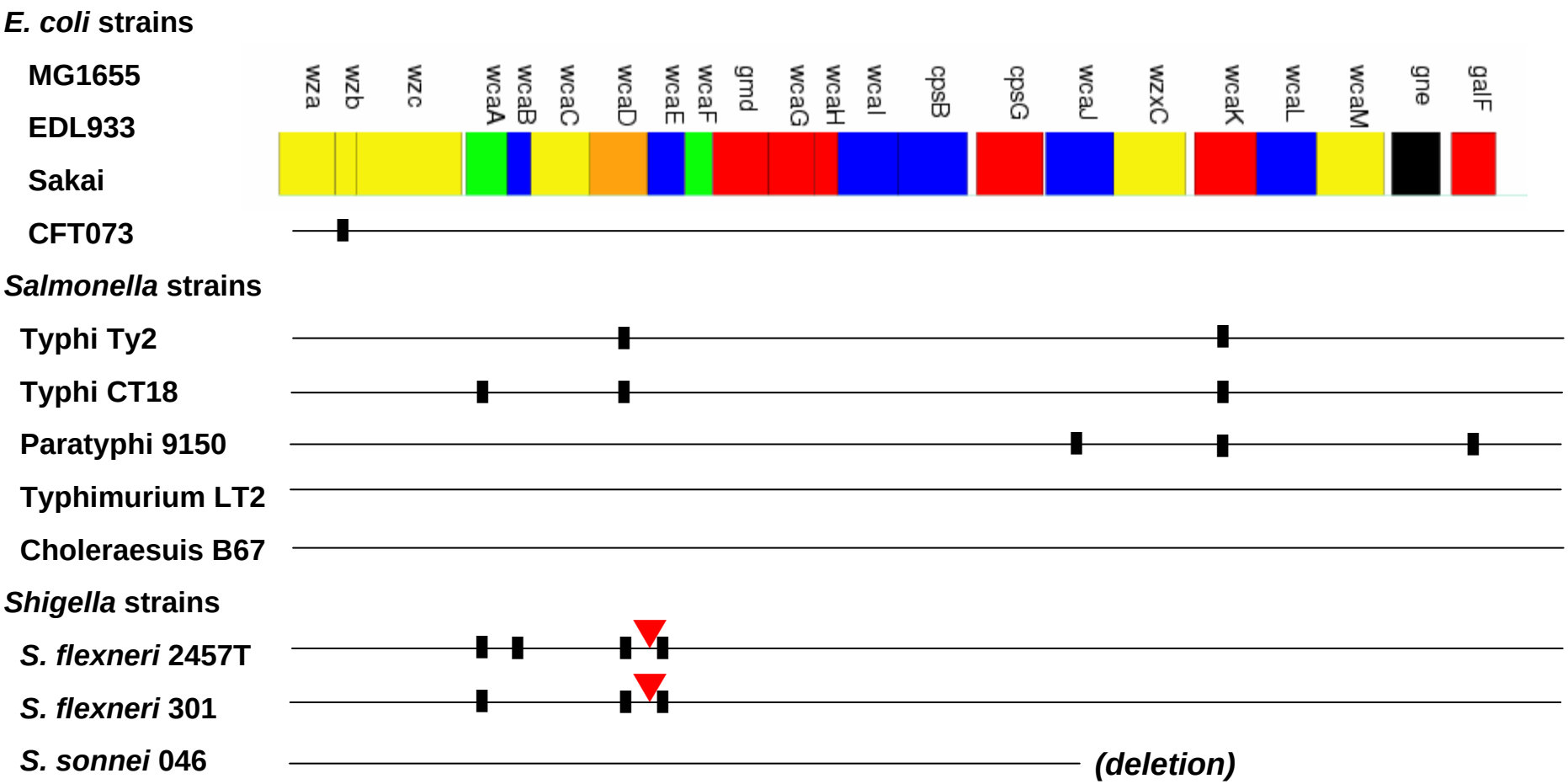
• In a given pathogen, are any **pseudogenes still expressed?** Is the **consequence** of a pseudogene something other than a straightforward loss of function? Altered enzyme activity, altered structural components, novel antigens …

• **Inconsistently annotated** in current genomes, making such questions harder to address.

Enteropathogen Resource Integration Center

Evolution, comparative genomics and genotyping: is a given gene present or absent in a particular genome, and is it intact or disrupted?

Example: pseudogenes in the colanic acid gene cluster



Enteropathogen Resource Integration Center

A possible issue for IOWG:

Do existing Sequence Ontology terms cover microbial pseudogenes?

SO:0000336 pseudogene

A sequence that closely resembles a known functional gene, at another locus within a genome, that is non-functional as a consequence of (usually several) mutations that prevent either its transcription or translation (or both). In general, pseudogenes result from either reverse transcription of a transcript of their "normal" paralog (SO:0000043) (in which case the pseudogene typically lacks introns and includes a poly(A) tail) or from recombination (SO:0000044) (in which case the pseudogene is typically a tandem duplication of its "normal" paralog).

SO:0000368 transposable_element_insertion_site

The junction in a genome where a transposable_element has inserted.

SO:0000718 blocked_reading_frame

A reading_frame that is interupted by one or more stop codons; usually identified through intergenomic sequence comparisons.

http://www.ericbrc.org

info@ericbrc.org

Updated Overview of ERIC Milestones and Timelines

ERIC	Q1	Q2	Q3	Q4
Year 1	▪ Staff Project ▪ Re-architect system ▪ Budget ▪ Software Development Planning ▪ On-going curation activities	▪ Order hardware ▪ IPP deliverables ▪ PDP Deliverable ▪ SWG Approved ▪ BRC Programmatic Meeting On-going curation activities	▪ Install system ▪ Deploy Initial Portal ▪ First SWG In-person Meeting ▪ On-going curation activities	▪ Obtain mAdb snapshot ▪ Mauve Integration ▪ Outreach- ASM, ISMB ▪ First SWG Meeting ▪ On-going curation activities
Year 2	▪ Preliminary ASAP integration completed ▪ Preliminary Synchronization development completed ▪ mAdb schema ported ▪ SWG Meeting at UW ▪ Outreach – Microbial Genomes ▪ On-going curation activities	▪ ASAP/ Synchronization Beta testing ▪ Preliminary mAdb integration completed ▪ Portal Software Change ▪ Outreach – US-Japan Panel, FWDIRN ▪ On-going curation activities	▪ ASAP integration ▪ Comparative genomics tools ▪ GBrowse ▪ mAdb Beta Testing ▪ Outreach- Mobile Elements, ASM Biodefense ▪ On-going curation activities	▪ Antigen/epitope tools ▪ Subsystem Integration ▪ SNP/Polymorphism Tools; Array Genotyping ▪ Continue importing databases into warehouse ▪ Text extraction ▪ Outreach – ASM, Beyond Genome - Bioinformatics ▪ On-going curation activities
Year 3	▪ Protein motif tools ▪ Pathway databases with graphic visualization ▪ Gene Prediction Tools ▪ Cross-platform tool (GoMiner, MatchMiner) integration ▪ On-going curation activities		▪ Transcription element prediction tools ▪ Proteomics and mass spectrometry comparison tools and data structures ▪ Pathogen/virulence/phenotype data structures ▪ Molecular interaction and visualization tools ▪ On-going curation activities	
Year 4	▪ Forensic tools ▪ Protein structural data modeling ▪ Structural data 3-D visualization tools ▪ On-going curation activities		▪ Biochemical data ▪ Protein structure prediction tools ▪ On-going curation activities	
Year 5	▪ Host-pathogen data structures ▪ Epidemiology ▪ On-going curation activities		▪ Data Mining Applications ▪ On-going curation activities	